文章编号:1673-9469(2012)02-0068-03

一种网络课程答疑系统分词器的设计

李龙[1],李丽丽[1],高玲[2]
(1.东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318;2.大庆油田图书馆,黑龙江 大庆 163453)

摘要:针对网络课程答疑系统提出了一种新的分词词典和查询算法,借鉴了现有三类分词算法的优点,克服了它们的不足,所设计的分词词典包括专业词典和基础词典两部分,所设计的算法在分词词典中搜索时,先搜索基础词典,后搜索专业词典,如果在基础词典中搜索出单词,则不继续搜索专业词典,该算法大大降低了算法的时间复杂度。本文将分词词典设计成由首字和次字构成的二维索引矩阵,和全部词语的有序顺序表组成,将单字的内码作为其在矩阵中的下标,对有序顺序表采用顺序查找,减少了词典搜索次数。

关键词:自然语言处理;答疑系统;分词;网络课程
中图分类号:TP182 **文献标识码**:A

Design of word segmentation implement in question - answering system of web - based course

LI Long[1], LI Li - li[1], GAO Ling[2]
(1. Computer and Information Technology College, NorthEast Petroleum University, Heilongjiang Daqing 163318, China; 2. Library of Daqing Oil Field, Heilongjiang Daqing 163453, China)

Abstract: This paper put forward a new word dictionary and query algorithms for network courses answering system, learn from the existing three types of sub - word segmentation algorithm of the advantages and overcome their shortcomings, the sub - word dictionary consists of two parts which are professional dictionary and the basic dictionary, the designed algorithm in the word dictionary search, searches firstly for the basic dictionary, searches secondly for the professional dictionaries. If finds words in the basic dictionary, the algorithm do not continue to search for specialized dictionaries, the algorithm greatly reduces the time complexity. In addition, word dictionary is designed by two - dimensional index of words and word matrix, and a table of all the ordered sequence of words, word within the code as a subscript in the matrix, the ordered sequence table using a sequential search which could reduce the number of dictionary search.

Key words: natural language processing; question - answering system; word segmentation; web - based course

在互联网技术广泛应用的今天,传统教学的弊端越来越凸显。将网络技术应用到教学,不仅能降低学习的门槛[1],还能突破时空的限制[2-5],提高教学效果。答疑系统作为网络课程帮助学生解答疑惑的平台,它替代传统答疑中教师的角色,直接与学生交流[6]。根据该原理,出现了一些基于WEB的答疑系统的研究[7-10]。但是,由于计算机很难真正理解学生提交的问题的含义,因此问题-解答库中即使有该问题的答案,也往往找不出来。针对此问题,文献[11]提出了对需求进行智能化展示的方法,文献[12]提出了基于所提问题与题库问题进行相似度计算的办法,在一定程

收稿日期:2012-03-10
作者简介:李龙(1967-),男,山东莱阳人,博士研究生,副教授,从事人工智能应用及石油工程计算技术。

度上改进了答案匹配的效果。相信相似匹配问题库中的问题和学生提交的问题切成一个个词语,然后将切分后的词语作为最基本的单元,执行相应的算法。该算法的前提是分词。现有的分词算法可分为3大类[13],分词词典的设计与查询算法是一大关键。现有的分词词典设计方法有[14]:基于整词二分的分词词典机制、基于Trie索引树的分词词典机制[15]、基于逐字二分的分词词典机制。基于整词二分的词典机制速度较慢,基于Trie索引树的分词词典机制,速度较快,但词典结构复杂,难以维护。第3种方法是前两种机制的折中,匹配效率提高有限,为了最大限度提高匹配效率,本文设计一种新的分词词典和查询算法,来满足网络课程答疑系统的需要。

1 分词词典设计

1.1 内容设计

对于网络课程答疑系统,用户输入的语句中非常有可能出现一些频率较低的专业单词,但绝大部分是使用频率较高的常用单词。因此,分词词典应该对应包括专业词典和基础词典两部分。

1.2 结构设计

将基础词典和专业词典中单词均按拼音排序,则排列后遵循如下:每个词典中的单词均按首字不同排列,首字相同的单词按次字不同排列,次字相同的单词按第三个字不同进行排列,依次类推。于是,首字相同的单词必定连续排在一起,首字相同且次字相同的单词也必定连续排在一起。因此,考虑采用Hash方法设计分词词典数据结构(见图1)。其中,数据区域采用顺序表存储所有单词,单词按拼音排序;索引区域由一个二维矩阵构成,包含特定首字和特定次字在数据区域存储的起始位置信息,行对应首字,列对应次字。

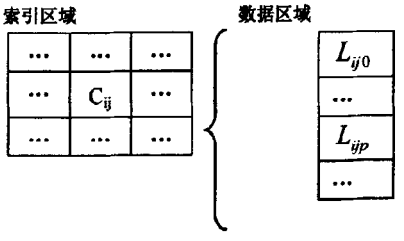

图1 分词词典数据结构

Fig. 1 Sub-word dictionary data structure

L_{ij0}至L_{ijp}是数据区域中首字为第i个字、次字为第j个字的全部单词。C_{ij}是首字为汉字中第i个字,次字为汉字中第j个字组成的单词对应的索引,其结构形式包括BeginPos和EndPos。

其中,BeginPos指示全部首字为第i个字、次字为第j个字组成的单词在数据区域中开始存储位置,EndPos指示在数据区域中结束存储位置。

1.3 索引矩阵确定

由于索引矩阵每一维均关联所有的汉字,因此它应该是一个方阵,且每一维的长度为汉字的个数。汉字的具体个数目前尚不清楚,一种说法是90 000多个,而《信息交换用汉字编码字符集基本集》GB2312-80中收录汉字6 763个,以此为标准构造分词词典,可以取索引矩阵为6 763*6 763的方阵。

在《信息交换用汉字编码字符集基本集》中,每个汉字均唯一对应一个内码。而根据转换公式:内码高位=区码+A0H(H为十六进制),内码低位=位码+A0H,可以很容易计算出一个汉字的区位码,再将区位码转换为十进制,即将最终得到的十进制数字作为该汉字在矩阵中的下标。用单词首字内码转换得到的十进制数作为它在索引矩阵中第一维的下标,用次字内码转换得到的十进制数作为它在索引矩阵中第二维的下标。

2 查询算法设计

由于语句中常用单词出现的概率,远比专业单词出现的概率大,因此,在分词词典中搜索时,应该先搜索基础词典,后搜索专业词典,如果在基础词典中搜索出单词,则不继续搜索专业词典。

2.1 算法描述

无论是专业词典,还是基础词典,均采用如下算法进行搜索。

定义首子串:对于汉字字符串T的长度不小于S的长度,若T和S从第一个字符一直到S的最后一个字符都相同,则称S是T的首子串。

算法描述:设输入字符串$T=(a_1a_2\cdots a_i\cdots a_n)$,其中$n$为字符串的长度($n>1$),$a_i$($i=1,2,\cdots,n$)表示字符串$T$的第$i$个字符。

计算汉字a_1的内码i,计算汉字a_2的内码j,获得词典索引矩阵中的第i行第j列的元素C_{ij}。

如果 C_{ij} 的第一个字段 BeginPos 的值 $\neq -1$,进入(3),否则说明词典中不存在单词(a_1),将(a_1)加入到临时集合 M 中,进入(5)。

如果 C_{ij} 的第二个字段 *EndPos* 的值 $\neq -1$,令 k = *Beginpos*,进入(4),否则说明词典中存在(a_1)这个单词,但不存在(a_1a_2)这个单词,将(a_1)加入到临时集合 M 中,进入(5)。

如果 $L[k]$ 是 T 的首子串,则字符串 T 中包含单词 $L[k]$,将 $L[k]$ 加入到临时集合 M 中,$k = k + 1\ 200$。若 $k >$ *EndPos*,则转至(6),否则转至(4)

将 M 中长度最大的词语作为搜索结果,结束搜索。

3 实验结果与分析

用 VS. NET 2005 实现了两个中文分词器。一个是采用基于整词二分的分词词典机制,另一个是采用上述方法。词典中的词条采用《汉语宝典》中的全部词语,并在 Pentium 43.0,1 024M 内存的情况下,进行了实验,实验结果如表 1。

表 1 系统测试结果

Tab. 1 System test results

词典机制	整词二分法	本文方法
时间/ms	186 455	752

从实验结果看,词典结构及相应查询机制对单词查询时间有很大影响。不采用索引逐字匹配,速度将会很慢。二者时间比为:186 455/752≈247.94

从理论上,二者时间相差的应该更多,本文的方法应该更快。

《汉语宝典》共收录双字词语 381 290 条,收录汉字 20 973 个[6]。采用整词二分法查找一个特定的多字词,从第一个单词到最后一个单词依次进行比较,假设每个单字出现的概率相等,则查找首字需要比较的次数为 $\frac{1}{381\ 290}\sum_{i=1}^{381\ 290} i = 190\ 645.5$,而平均包含特定首字的词语有 381 290/20 973≈18 个,即查找该首字之后不同的次字平均要比较 18 次。因此,查找首字和次字平均需要比较 190 645.5 + 18 = 190 663.5 次。采用本文词典结构和搜索方法,查找某特定词语的首字和次字只需要计算汉字内码 2 次。二者查询次数比为:90 663.5/2 = 95 331.75。

可以看出,理论时间比为 95 331.75,实际时间比为 247.94。通过分析程序,得知两种算法都需要加载词库,加载词库需要耗费大量时间,该时间是算作总时间之内的,因此实际时间比减小了。

4 结束语

本文所给出的一种新的词典设计方法和查询算法,大大降低了算法的时间复杂度。但算法的实际运行时间并没有如时间复杂度预计减少的那么多,但可以研究词库加载算法,进一步提高词语的实际查询速度。

参考文献:

[1] 胡青松,张 申. 通用网络辅助教学支撑平台的研制[J]. 电气电子教学学报,2008(03):74-76.

[2] 张 玮. 基于网络交互的学习共同体研究[J]. 软件导刊(教育技术),2011(09):25-28.

[3] 桑新民. 现在教育技术学基础理论创新研究[J]. 中国电化教育,2003(9):56-59.

[4] 韩海英. 基于网络化教学环境的教师角色重塑[J]. 教育革新,2009(01):21-22.

[5] 姜大仲,王新秀,崔善珠. 发展终身学习型城市网络的战略[J]. 高等函授学报:哲学社会科学版,2011(05):3-6.

[6] 武法提. 网络教育应用[M]. 北京:高等教育出版社,2003.

[7] 姜良华. 网络辅助答疑系统的设计与实现[J]. 电脑知识与技术,2011(26):6451-6452.

[8] 方光伟. 基于 Web 的课程自动答疑系统的设计与实现[J]. 科技信息,2011(16):197-198.

[9] 王 薇,朱 凤,李 欢. 基于 Web 的课程答疑系统的研究[J]. 中国成人教育,2008(22):159-160.

[10] 蔡冠群,张业睿,袁晓斌. 构筑基于 Web 的远程答疑系统[J]. 信息技术教育,2006(03):75-76.

[11] 朱云霞,周海峰. 基于 WEB 的智能答疑系统的研究与设计[J]. 科技信息,2009(01):413-414.

[12] 康文宁,杨志强. 相似度计算在智能答疑系统中的研究及应用[J]. 计算机技术与发展,2010(2):71-74.

[13] 文庭孝. 汉语自动分词研究进展. 图书情报[J],2005(5):54-62.

[14] YOU C H, KOH S N, RAHARDJA S. An invertible frequency eigendomain transformation for masking-based subspace speech enhancement[J]. IEEE Signal Processing Letters,2005,12(6):461-464.

[15] 严蔚敏,吴伟民. 数据结构(C 语言版)[M]. 北京:清华大学出版社,2003.

(责任编辑 刘存英)